In the Claims:

This section sets forth a clean version of the entire set of pending claim(s) under 37 C.F.R. 1.121(c)(3) and MPEP 714.22(a). Appendix A submitted herewith sets forth a marked up version of the prior pending claim(s) which have been amended by this Amendment with additions shown with underlining (e.g. new text) and deletions shown with a strikethrough (e.g. ~~delete text~~) under 37 C.F.R. 1.121(c)(1)(ii).

---

1. A method for quantitatively representing digital documents in a vector space, comprising the steps of:

identifying a first digital document to be processed from a plurality of digital documents;

extracting a first feature corresponding to the first document from the plurality of digital documents, the first feature comprising text surrounding an image included in the digital document, the text surrounding the image not being anchor text;

converting the first feature to a first vector; and

associating the first vector with the first digital document.

2. ~~1.~~ The method of claim 1 further comprising the steps of:

extracting a second feature corresponding to the digital document, the second feature comprising a first URL representing the first digital document;

converting the second feature to a second vector; and

associating the second vector with the first digital document.

3. ~~8.~~ The method of claim ~~7~~ 2, wherein the step of converting the second feature comprises the sub-steps of:

identifying each unique word within the URLs representing all digital documents in the collection of digital documents; and

counting the occurrences of each unique word in the first URL;

creating a vector having a number of dimensions equal to the number of unique words in the URLs representing all digital documents in the collection of digital

documents, and further having as each element a numeric value representative of the number of occurrences in the first URL of the corresponding word.

~~8.~~ The method of claim ~~8~~, wherein the value representative of the number of occurrences in the first URL of the corresponding word is calculated as the token frequency weight of the corresponding word multiplied by the inverse context frequency weight of the corresponding word.

~~10.~~ The method of claim 1 further comprising the steps of:

extracting a second feature corresponding to the first digital document, the second feature comprising inlinks in the collection of digital documents linking to the first document;

converting the second feature to a second vector; and

associating the second vector with the first digital document .

~~11.~~ The method of claim ~~10~~, wherein the step of converting the second feature comprises the sub-steps of:

identifying each digital document having links within the collection of digital documents;

determining how many times each digital document having links points to the first digital document; and

creating the second vector having a number of dimensions equal to the number of digital documents having links in the collection of digital documents, and the second vector further having as each element a numeric value representative of the number of links in each corresponding digital document linking to the first digital document.

~~12.~~ The method of claim ~~11~~, wherein the numeric value representative of the number of links in each corresponding digital document linking to the first digital document is calculated as the token frequency weight of the corresponding link multiplied by the inverse context frequency weight of the corresponding link.

8 ~~13.~~ The method of claim 5 ~~10~~, wherein the step of converting the second feature comprises the sub-steps of:

identifying each digital document having hyperlinks within the collection of digital documents, and further identifying each unique word associated with URLs defining hyperlinks in each digital document;

counting the occurrences of each unique word in the URLs defining hyperlinks pointing to the first digital document; and

creating the second vector having a number of dimensions equal to the number of unique words associated with URLs defining hyperlinks within the collection of digital documents, and the second vector further having as each element a numeric value representative of the number of occurrences in the URLs defining hyperlinks pointing to the first digital document of the corresponding word.

9 ~~14.~~ The method of claim 8 ~~13~~, wherein the numeric value representative of the number of occurrences in the URLs defining hyperlinks pointing to the first digital document of the corresponding word is calculated as the token frequency weight of the corresponding word multiplied by the inverse context frequency weight of the corresponding word.

10 ~~15.~~ The method of claim 1 further comprising the steps of:

extracting a second feature corresponding to the first digital document, the second feature comprising outlinks in the collection of digital documents linking to the first digital document;

converting the second feature to a second vector; and

associating the second vector with the first digital document.

11 ~~16.~~ The method of claim 10 ~~15~~, wherein the step of converting the second feature comprises the sub-steps of:

identifying each other digital document linked to by all digital documents within the collection of digital documents; and

creating the second vector having a number of dimensions equal to the number of other digital documents linked to by digital documents in the collection of digital documents, and the second vector further having as each element a numeric value representative of the number of links in the first digital document linking to each corresponding other digital document.

12
~~17~~. The method of claim ~~16~~ 11, wherein the numeric value representative of the number of links in the first digital document linking to each corresponding other digital document is calculated as the token frequency weight of the corresponding link multiplied by the inverse context frequency weight of the corresponding link.

13
~~18~~. The method of claim ~~15~~ 10, wherein the step of converting the second feature comprises the sub-steps of:

identifying each unique word associated with URLs defining hyperlinks in each digital document in the collection of digital documents;

counting the occurrences of each unique word in the URLs defining hyperlinks in the first digital document; and

creating the second vector having a number of dimensions equal to the number of unique words associated with the URLs defining hyperlinks in each digital document, and the second vector further having as each element a numeric value representative of the number of occurrences in the URLs defining hyperlinks in the first digital document of the corresponding word.

14
~~19~~. The method of claim ~~16~~ 13, wherein the numeric value representative of the number of occurrences in the URLs defining hyperlinks in the first digital document of the corresponding word is calculated as the token frequency weight of the corresponding word multiplied by the inverse context frequency weight of the corresponding word.

24
~~20~~. The method of claim ~~19~~ 23, wherein the second feature comprises a text genre feature.

25 24

21. The method of claim 20, wherein the step of converting the second feature comprises the sub-steps of:

for each possible text genre, processing the first digital document to calculate the probability that the first digital document is of the corresponding text genre; and

creating the second vector having a number of dimensions equal to the number of possible text genres, and the second vector further having as each element a numeric value representative of the probability that the first digital document is of the corresponding genre.

24 23

22. The method of claim 19, wherein the first feature comprises the color histogram for the image included in the first digital document.

27 26

23. The method of claim 22, wherein the step of converting the first feature comprises the sub-steps of:

quantizing the image represented by the first digital document into a multi-dimensional color model;

creating a color histogram having a plurality of bins for each dimension in the color model, each bin corresponding to a unique combination of binary bits representing information from the associated dimension of the color model;

counting each of a plurality of pixels from the image in a corresponding bin associated with each dimension of the color model; and

creating the first vector having a number of dimensions equal to the total number of bins in the color histogram, and the first vector further having as each element a numeric value representative of the number of pixels in the image corresponding to the corresponding histogram bin.

28 27

24. The method of claim 23, wherein the plurality of pixels from the image in the counting step comprises all of the pixels in the image.

29 28
25. The method of claim 24, wherein the plurality of pixels from the image in the counting step comprises an approximately uniformly spaced set of subsampled pixels from the image.

30 27
26. The method of claim 23, wherein:

the color model comprises a three-dimensional hue, saturation, and value color model;

each dimension of the color model is represented by two bits of information; and

the color histogram has four bins for each dimension in the color model, for a total of twelve bins.

31 27
27. The method of claim 23, wherein the image represented by the first digital document comprises a region of a bitmap.

32 23
28. The method of claim 19, wherein the first feature comprises the color complexity of an image represented by the first digital document.

33 32
29. The method of claim 28, wherein the step of converting the first feature comprises the sub-steps of:

quantizing the image represented by the first digital document into a multi-dimensional color model;

determining the maximum number of pixels in any row in any image represented by any digital document in the collection of digital documents;

determining the maximum number of pixels in any column in any image represented by any digital document in the collection of digital documents;

creating a horizontal complexity histogram and a vertical complexity histogram, each having a number of bins equal to the maximum number of pixels in any row and in any column, respectively;

identifying horizontal runs of pixels of all possible lengths in the quantized image, and for each possible length, counting the number of pixels in a plurality of rows of the

quantized image belonging to the horizontal runs in a corresponding bin of the horizontal complexity histogram;

identifying vertical runs of pixels of all possible lengths in the quantized image, and for each possible length, counting the number of pixels in a plurality of columns of the quantized image belonging to the vertical runs in a corresponding bin of the horizontal complexity histogram;

creating a horizontal complexity vector having a number of dimensions equal to the maximum number of pixels in any row, and further having as each element a numeric value representing the number of pixels in the image in the corresponding horizontal histogram bin; and

creating a vertical complexity vector having a number of dimensions equal to the maximum number of pixels in any column, and further having as each element a numeric value representing the number of pixels in the image in the corresponding vertical histogram bin.

34 33
30. The method of claim ~~29~~, wherein the plurality of rows comprises all rows of the quantized image, and wherein the plurality of columns comprises all columns of the quantized image.

35 33
~~31~~. The method of claim ~~29~~, wherein the plurality of rows comprises an approximately uniformly spaced set of subsampled rows from the image, and wherein the plurality of columns comprises an approximately uniformly spaced set of subsampled columns from the image.

36 33
~~32~~. The method of claim ~~29~~, wherein:

the color model comprises a three-dimensional hue, saturation, and value color model; and

each dimension of the color model is represented by two bits of information.

37 33

~~33.~~ The method of claim ~~29~~, further comprising the step of concatenating the horizontal complexity vector and the vertical complexity vector to form a complexity vector having a number of dimensions equal to the maximum number of pixels in any row plus the maximum number of pixels in any column.

38 32

~~34.~~ The method of claim ~~28~~, wherein the step of converting the first feature comprises the sub-steps of:

quantizing the image represented by the first digital document into a multi-dimensional color model;

determining the maximum number of pixels in any row in any image represented by any digital document in the collection of digital documents;

determining the maximum number of pixels in any column in any image represented by any digital document in the collection of digital documents;

creating a horizontal complexity histogram and a vertical complexity histogram, each having a selected number of bins corresponding to a plurality of quantized ranges of run lengths;

identifying horizontal runs of pixels of all possible lengths in the quantized image, and for each possible length, counting the number of pixels in a plurality of rows of the quantized image belonging to the horizontal runs in a corresponding bin of the horizontal complexity histogram;

identifying vertical runs of pixels of all possible lengths in the quantized image, and for each possible length, counting the number of pixels in a plurality of columns of the quantized image belonging to the vertical runs in a corresponding bin of the horizontal complexity histogram;

creating a horizontal complexity vector having a number of dimensions equal to the selected number of bins in the horizontal complexity histogram, and further having as each element a numeric value representing the number of pixels in the image in the corresponding horizontal histogram bin; and

creating a vertical complexity vector having a number of dimensions equal to the number of bins in the vertical complexity histogram, and further having as each element

a numeric value representing the number of pixels in the image in the corresponding vertical histogram bin.

39 38
~~35~~. The method of claim ~~34~~, wherein:

a bin b_x in the horizontal complexity histogram corresponding to a horizontal run of length r_x is identified by a relationship $b_x = \text{floor}(r_x(N\text{-}1) / (n_x/4)) + 1$, where N is the selected number of bins in the horizontal complexity histogram and n_x is a maximum number of pixels in any row of an image in the collection; and

a bin b_y in the vertical complexity histogram corresponding to a vertical run of length r_y is identified by a relationship $b_y = \text{floor}(r_y(N\text{-}1) / (n_y/4)) + 1$, where N is the selected number of bins in the horizontal complexity histogram and n_y is a maximum number of pixels in any row of an image in the collection.

40 38
~~36~~. The method of claim ~~34~~, wherein the plurality of rows comprises an approximately uniformly spaced set of subsampled rows from the image, and wherein the plurality of columns comprises an approximately uniformly spaced set of subsampled columns from the image.

41 38
~~37~~. The method of claim ~~34~~, wherein:

the color model comprises a three-dimensional hue, saturation, and value color model; and

each dimension of the color model is represented by two bits of information.

42 38
~~38~~. The method of claim ~~34~~, further comprising the step of concatenating the horizontal complexity vector and the vertical complexity vector to form a complexity vector having a number of dimensions equal to the selected number of bins in the horizontal complexity histogram plus the selected number of bins in the vertical complexity histogram.

15
~~39.~~ A signal representing instructions for quantitatively representing in a vector space users of a collection of digital documents, the instructions comprising:

identifying a first user to be processed from the users of the collection of digital documents;

extracting from the collection of digital documents a first feature representing a first sub-set of digital documents of the collection that have been accessed by the first user;

converting the first feature to a first vector; and

associating the first vector with the first user.

16
~~41.~~ The signal of claim 15, ~~39~~, wherein the converting instruction comprises:

identifying each unique digital document in the collection of digital documents;

calculating the number of times the first user accessed each digital document in the collection of digital documents; and

creating the first vector having a number of dimensions equal to the number of digital documents in the collection of digital documents, and the first vector further having as each element a numeric value representative of the number of times the first user has accessed the corresponding digital document.

17
~~42.~~ The signal of claim 16, ~~41~~, wherein the value representative of the number of times the first user has accessed the corresponding digital document is calculated as the token frequency weight of the corresponding digital document multiplied by the inverse context frequency weight of the corresponding digital document.

18
~~43.~~ A computer-readable medium containing instructions for causing a computer-system to quantitatively represent digital documents in a vector space, by the steps of:

identifying a digital document to be processed from a plurality of digital documents;

selecting an image feature as a first feature, the image feature being associated with the non-text content of an image included in the digital document;

extracting from the document information associated with the first feature;

converting information associated with the first feature into a first vector;

associating the first vector with the digital document;

selecting a second feature from a set of multi-modal features including a user information feature and a genre feature;

extracting from the document information associated with the second feature;

converting the information associated with the second feature into a second vector; and

associating the second vector with the digital document.

19 18

~~45.~~ The computer-readable medium of claim ~~43~~ wherein the first feature comprises a color histogram for the image included in the digital document.

20 19

~~46.~~ The computer-readable medium of claim ~~45~~ wherein converting the information associated with the first feature into the first vector comprises the steps of:

quantizing the image included in the digital document into a multi-dimensional color model;

creating a color histogram having a plurality of bins for each dimension in the color model, each bin corresponding to a unique combination of binary bits representing information from the associated dimension of the color model;

counting each of a plurality of pixels from the image in a corresponding bin associated with each dimension of the color model; and

creating a vector having a number of dimensions equal to the total number of bins in the color histogram, and further having as each element a numeric value representative of the number of pixels in the image corresponding to the corresponding histogram bin.

21 18
~~47.~~ The computer-readable medium of claim ~~43~~ wherein the first feature comprises color complexity of the image included in the digital document.

22 21
~~48.~~ The computer-readable medium of claim ~~47~~ wherein converting the information associated with the first feature into the first vector comprises the steps of:

quantizing the image included in the digital document into a multi-dimensional color model;

determining the maximum number of pixels in any row in any image represented by any digital document in the collection of digital documents;

determining the maximum number of pixels in any column in any image represented by any digital document in the collection of digital documents;

creating a horizontal complexity histogram and a vertical complexity histogram, each having a number of bins equal to the maximum number of pixels in any row and in any column, respectively;

identifying horizontal runs of pixels of all possible lengths in the quantized image, and for each possible length, counting the number of pixels in a plurality of rows of the quantized image belonging to the horizontal runs in a corresponding bin of the horizontal complexity histogram;

identifying vertical runs of pixels of all possible lengths in the quantized image, and for each possible length, counting the number of pixels in a plurality of columns of the quantized image belonging to the vertical runs in a corresponding bin of the horizontal complexity histogram;

creating a horizontal complexity vector having a number of dimensions equal to the maximum number of pixels in any row, and further having as each element a numeric value representing the number of pixels in the image in the corresponding horizontal histogram bin; and

creating a vertical complexity vector having a number of dimensions equal to the maximum number of pixels in any column, and further having as each element a numeric value representing the number of pixels in the image in the corresponding vertical histogram bin.

23

~~40~~. A method for quantitatively representing digital documents in a vector space, comprising the steps of:

identifying a first digital document to be processed from a plurality of digital documents;

extracting a first feature corresponding to the first digital document from the plurality of digital documents, the first feature comprising an image feature associated with non-text content of an image included in the first digital document;

converting the first feature to a first vector;

associating the first vector with the first digital document;

extracting a second feature corresponding to the digital document, the second feature comprising a one of a user feature and a text genre feature;

converting the second feature into a second vector; and

associating the second vector with the first digital document.